1 Related Work

Since the 2016 U.S. presidential election, fake news has been a frequent topic of natural language processing research. There are countless examples of papers approaching fake news classification or closely related problems, but from slightly different angles. This section summarizes the prior work in fake news detection, clarifying the different categories of methods. In general, previous works in fake news detection differ in three major ways: 1) the scale of the predicted variable, the information used as features, and the type of model. With respect to scale, any fake news detection model falls into one of three levels of granularity: 1) claim level, 2) source level, and 3) article level. These levels of analysis describe the response variable being predicted.

Claim level approaches attempt to determine whether a given claim, usually one or several sentences, is true or intentionally misleading. [1] Given that claim-level approaches must make a judgement based on only a short amount of text, researchers often adopt a fact-checking strategy This strategy, also known as "truth discovery," assumes that a sentence's claims can be gramatically isolated and checked against a database of established claims.[2] A natural application for claim-level models is social media, most commonly Twitter, where little is known about the author and only a very limited amount of text is available. However, claim-level approaches have serious limitations, often struggling with the complex sentences journalists or other writers typically employ.[3] Aditionally, they rely entirely on a complete knowledge base, which must be constantly expanded and updated, clearly a difficult task.

Source level approaches attempt to classify whether a speaker or entire news source consitently publish misinformation. The intuition behind these approaches is that speakers or sources that have published misinformation in the past are likely to continue to do so. An example of a source-level approach is the popular browser extension "BS Detector,"[4] which classifies articles on a fine-grained scale of veracity by checking the source's status in a database of news sources and their reliability. A source's history of misinformation can also be used as a predictor in claim-level or article level approaches. Kirilin and Strube,

[1] Examples of claim-level approaches include ...

[2] strube p. 2

[3] strube 2

[4]

for instance, create *Speaker2Credit*, a metric of speaker credibility, and show how it can improve the performance of fake news detection models when used as an input.[5]

Article level approaches have received the most attention in the work on fake news detection. This is logical, given that fake news tends to take the form of articles, peddling misinformation in the article text while posing as a legitimate source. Article-level approaches also benefit from a rich list of predictors to choose from, including not only the article's content but all relevant metadata. Kai Shu et. al, for instance, build an article-level model using linguistic and visual components of the article content, the social context around it—including information on the user that posted it, the post itself, responses to it, and the social network of the poster—, as well as spatiotemporal information capturing when and where the article and responses were to it were posted from.[6]

The work of Shu et. al is an example of the overwhelming number predictors available to researchers working in fake news detection, resulting in a diversity of approaches with respect to feature selection. Melanie Tosik et. al, for instance, employ only hand-crafted features capturing the similarity between an article's title and text in a two-stage ensemble classifier modeling whether an article's body agrees with its headline.[7] Sonam Tripathi and Tripti Sharma demonstrate the effectiveness of parts of speech tagging—also known as grammatical tagging—in document classification problems, the general category of natural language processing that article-level fake news detection falls under.[8] Ramy Baly et al. employ a breadth of features to model factuality and bias of news sources, using features covering the content of articles, the source's Wikipedia and Twitter pages, the structure of the URL, and the source's web traffic. [9]

Independent of level of analysis or choice of predictors, researchers overwhelmingly choose to use complex deep neural networks. Ajao et. al, for instance, use a "hybrid of convolutional neural networks and long-short term recurrent neural network models" to classify Tweets as true or false based on their text content.[10] The dominance of deep learning approaches is visible in an extensive survey on fake news detection done by Ray Oshikawa and Jing Qian. [11] The pair's section on machine learning models dedicates a total of three sentences to "Non-Neural Network Models," compared to seven paragraphs. focusing on neural networks.

While deep learning approaches can produce highly accurate models that consistently succeed in identifying misinformation, they also suffer from a lack of interpretability. This

[5] Kirilin and Strube
[6] Shu et al
[7] Tosik et al
[8] tripathi
[9]
[10] ajao et al
[11] Oshikawa

is often referred to as the "black-box" problem, meaning that the inputs and outputs of these models are perfectly clear, but the steps that the model takes to reach the output are completely invisible. This has been identified as a limitation of the the work on fake news detection thus far.[12] Oshikawa and Qian, at the conclusion of their extensive survey, declare that "we need more logical explanation for fake news characteristics," highlighting the need for models that can teach us something about fake news.

Approaches that focus on interpretability are rare, but do exist. From the deep learning perspective, Nicole O'Brien et al. employ post-hoc variable importance to their text-based deep learning model, identifying the words that are most predictive of fake and real news.[13] Their approach, however, does not interpret the results, but instead merely demonstrates the feasiblity of the technique. Furthermore, this method reveals only information about *specific* words, as opposed to *types* of words. More applicable is the work of researchers who both use features that describe the semantic properties of the text in general, use interpretable models, and extensively document their results. The best examples of this type of work are the works of Benajmin Horne et al. and Mauricio Gruppi et al., both of which employ features capturing the complexity, style, and psychology of fake news, and display precisely how fake and real news differs in each of the variables used.[14]

[12] shu, O'Brien
[13] O'Brien et al
[14] Horne, Gruppi